CAAP Quarterly Report

06/29/2023

Project Name: Development of Compatibility Assessment Model for Existing Pipelines for Handling Hydrogen-Containing Natural Gas

Contract Number: 693JK32250004CAAP

Prime University:

Prepared By: Ramadan Ahmed, r.ahmed@ou.edu, 405-473-7771

Reporting Period: 03/30/2023 – 06/29/2023

Project Activities for Reporting Period:

Project team members continued to develop a database for hydrogen embrittlement of carbon steel used in pipeline applications (Task 1). The team also continued to modify experimental facilities so that the planned laboratory studies could be conducted (Task 2.1). The designs of test specimens for tensile strength (Task 2.2), fracture toughness (Task 2.3), and fatigue resistance (Task 2.4) have been finalized and sent to a local workshop for manufacturing. Due to manufacturing delays, we are still waiting for autoclaves. As indicated in the previous report, the start date for experimental studies on tensile properties (Task 2.2) will be postponed. We have ordered an additional autoclave to conduct tests in parallel to avoid delays in subsequent tasks (Tasks 2.2 to 2.4).

We have gathered experimental data sets concerning hydrogen diffusivity, solubility, area reduction, maximum elongation, fracture toughness, and fatigue resistance to develop a master database for pipeline materials' hydrogen embrittlement (HE). Afterward, imputation or removal techniques are applied to resolve incomplete data issues. Data from various sources are used when metal composition and mechanical characteristics are lacking. The data analysis task (Task 1.3) has been completed, and the results are summarized in Appendix A.

Project Financial Activities Incurred during the Reporting Period:

Table 1 presents expenses during the reporting period in each budget category.

Table 1: Quarterly expense breakdown

Budget Category	DOT-PHMSA	OU Cost Share	Total
Salaries and Wages	$30,417	$4,286	$34,703
Fringe Benefits	$6,207	$1,474	$7,681
Equipment	$2,684		$2,684
Travel			
Materials and Supplies			
Tuition	$10,933	$1,474	$12,407
Indirect Costs			
Total	**$50,241**	**$7,234**	**$57,475**

Project Activities with Cost Share Partners:

As part of the cost sharing, the Principal Investigator and Co-PI participated in various research and development activities, including supervising research assistants and technical personnel, conducting hydrogen embrittlement research, and designing experimental setups.

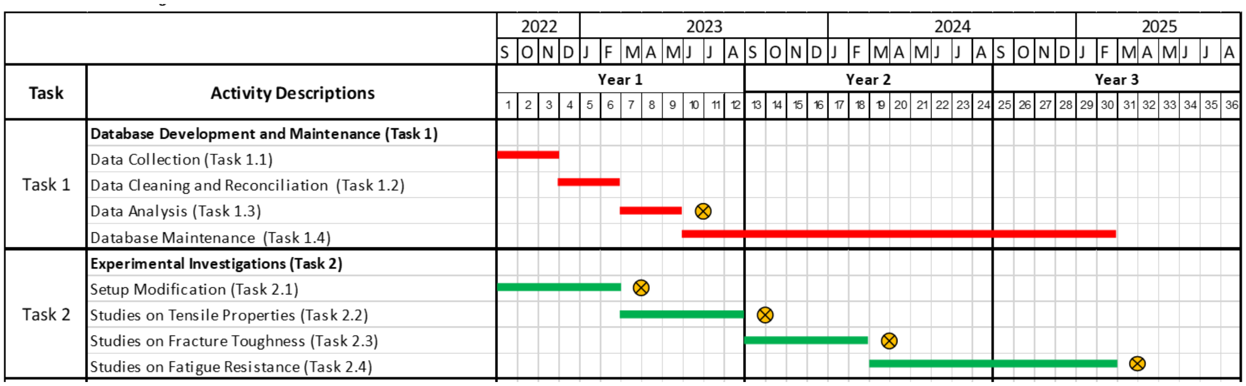
Project Activities with External Partners:

Not applicable.

Potential Project Risks:

We anticipate a four to five month delay in modifying experimental facilities (Task 2.1) due to autoclave manufacturing delays. Since experimental studies (Tasks 2.2 and 2.3) will take place in parallel, the delay won't have a significant impact on the overall project schedule (**Table 2**).

Table 2: Project schedule



Future Project Work:

In the coming weeks, we will receive the autoclaves and complete the modification of the experimental setup (Task 2.1). Currently, a local workshop manufactures all specimens (450 pieces) required for experimental investigations. Manufacturing will be completed in the coming weeks. We anticipate starting experiments in 3-4 weeks. Furthermore, we will prepare and submit papers for journal publications to present our findings from our data analysis in Task 1.3. The papers present results of data analysis and the development and testing of machine learning algorithms to predict fatigue crack growth rate and area reduction in hydrogen environments.

Potential Impacts to Pipeline Safety:

By using machine learning algorithms developed at the current phase of the project, we can forecast the level of degradation that will occur if hydrogen is transported through existing pipelines. Using these predictions, a safe hydrogen transportation envelope can be established in the existing natural gas pipeline systems.

Appendix: Summary of Data Analysis (Task 1.3)

Table 3 shows summary statistics for the transformed variables in the fatigue dataset prepared for data analysis. During data processing, a data quality report was created. The dataset showed incomplete values of Ultimate Tensile Stress (Sx). The predictive mean matching (PMM) method was used to forecast these missing values. Moreover, missing values for the Heat Treatment parameter were handled by creating another factor level called "Unknown" since they represent data points not reported in the original data sources.

During the data analysis, the distribution of each parameter is examined and compared with the distribution obtained by transforming the parameter using powers in the range of -3 to +3. Data transformation and normalization techniques are applied to address problems with abnormalities and disproportionate variances within variables. The methods are critical to having a dataset that generalizes as much as possible to avoid overfitting and performs well with unseen data. They could also help find relationships between variables that may not exist in their current forms but can be found in a different dimension. These methods also help address perceived outliers in a dataset. **Figure 1** shows the distribution of crack extension per load cycle (da/dN), stress intensity factor range (DK), H2 Pressure, and Yield Stress (Sy) in different powers or dimensions. A log transform of these four variables generates near-normal distributions compared with the original ones. (Power 1). Other numerical variables either exhibit near-normal distributions in their original forms or show no significant change in distribution in different dimensions. Therefore, they are not transformed.

Table 3: Summary statistics for variables in fatigue dataset

Variable	N	Mean	Std. Dev.	Min	25%	75%	Max
Metal Fe content	1360	98.466	0.291	97.8	98.3	98.706	99.312
Metal C content	1360	0.214	0.19	0.048	0.12	0.26	0.85
Metal Mn content	1360	1.121	0.284	0.47	0.82	1.29	1.53
Metal P content	1360	0.012	0.005	0	0.008	0.014	0.02
Metal S content	1360	0.019	0.011	0	0.012	0.026	0.042
Metal Si content	1360	0.163	0.098	0	0.11	0.25	0.31
Gas hydrogen pressure	1360	0.695	0.459	0	0	1	1
Gas CO2 content	1360	0.001	0.004	0	0	0	0.02
Gas SO2 content	1360	0.001	0.004	0	0	0	0.02
Gas O2 content (ppm)	1360	3.382	18.084	0	0	0	100
Frequency	1360	1.146	1.198	0.1	1	1	5
Load ratio	1360	0.258	0.235	0.05	0.1	0.5	0.8
ΔK	1360	2.936	0.453	1.716	2.635	3.255	4.538
Sy	1360	6.056	0.203	5.595	5.903	6.159	6.461
Sx	1360	610.29	87.287	463	526	675	835
da/dN	1360	-9.301	2.558	-20.592	-10.668	-7.566	-4.063

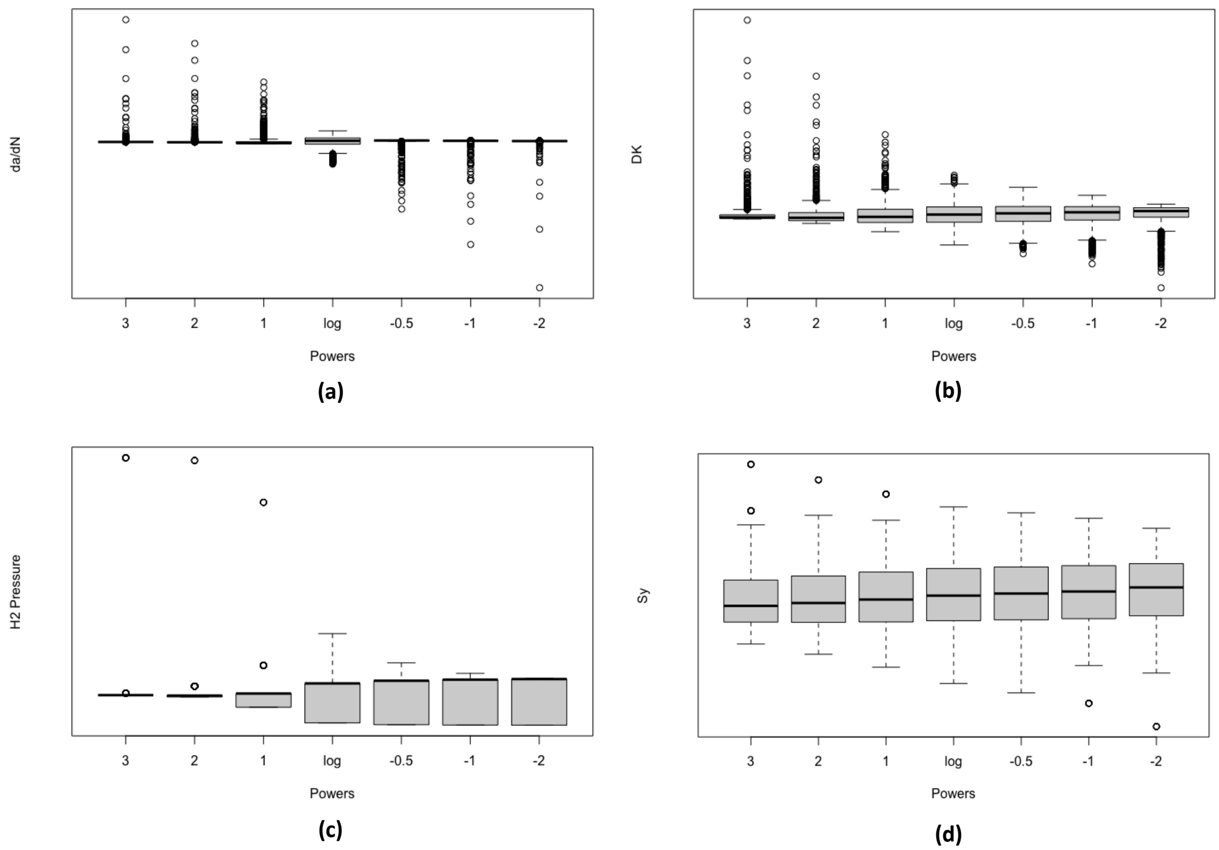


Fig. 1: Distribution of (a) da/dN, (b) ∆K, (c) H2 Pressure, and (d) Sy in different dimensions

The relationships between the numerical variables and crack extension per load cycle (da/dN) are examined by creating a cross-correlation plot (**Fig. 2**). It is found that there is a strong correlation between crack extension and stress intensity factor range (DK) and hydrogen pressure. It is easy to observe that steel composition (carbon, phosphorous, and sulfur concentrations), gas content (O2, SO2, and CO2 concentrations) and test parameters (load ratio and load frequency) all influence steel crack extension. As part of exploratory data analysis, distribution plots of log-transformed da/dN values obtained with and without hydrogen presence in the test environment are shown in **Fig 3**. The dashed lines represent the mean of the distributions. Hydrogen mediums result in higher fatigue crack growth rates (da/dN).

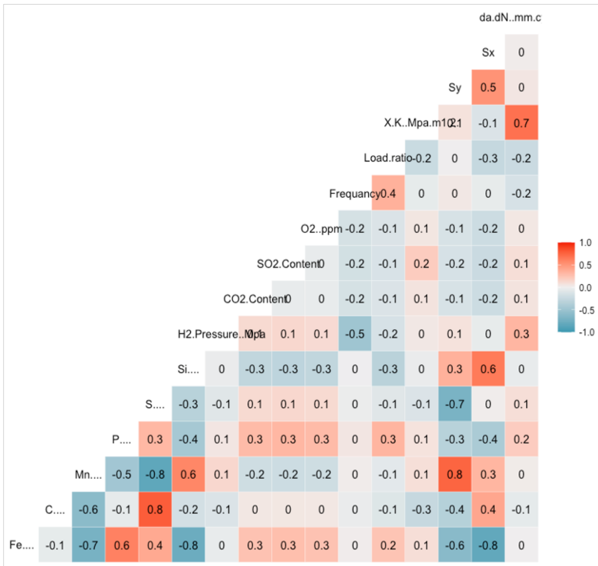


Fig. 2: Cross-correlation plot

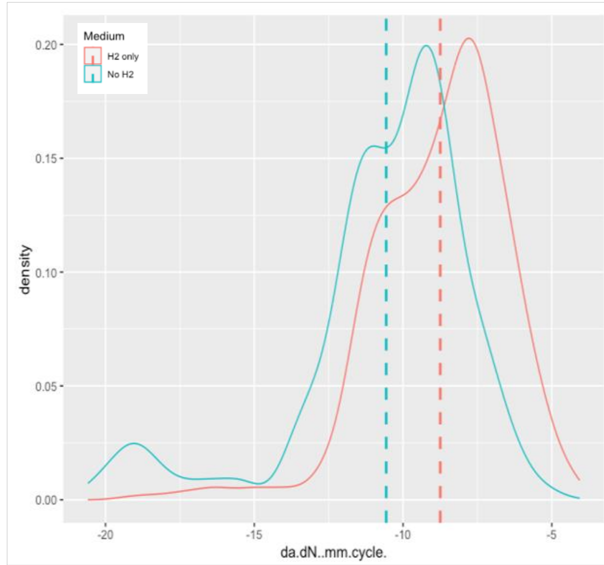


Fig. 3: da/dN distributions with and without hydrogen

In addition to data analysis, a preliminary machine learning (ML) modeling has been conducted. A multiple linear regression (linear regression) model, a multivariate adaptive regression spline (MARS) model, a random forest, and a gradient boosting method are selected for modeling. These models (algorithms) are chosen due to their ability to handle non-linear relationships between variables, adaptability to complex datasets, and proven performance in similar predictive tasks. **Figure 4** compares different models' Root Mean Square Error (RMSE). The Random Forest (RF) model has the lowest RMSE compared to other models. Gradient Boosting (XGBoost) exhibits the lowest RMSE on training data but does not perform as well on test data as Random Forest (RF). **Figure 5** compares the measured and predicted transformed values of da/dN using the Random Forest model on training and test data. The model is reasonably accurate in reproducing observed values in both cases. For some date points, slightly increased discrepancies are observed when predictions are compared with the test dataset at high transformed da/dN values. Model performance was reduced with training data at low transformed da/dN values.

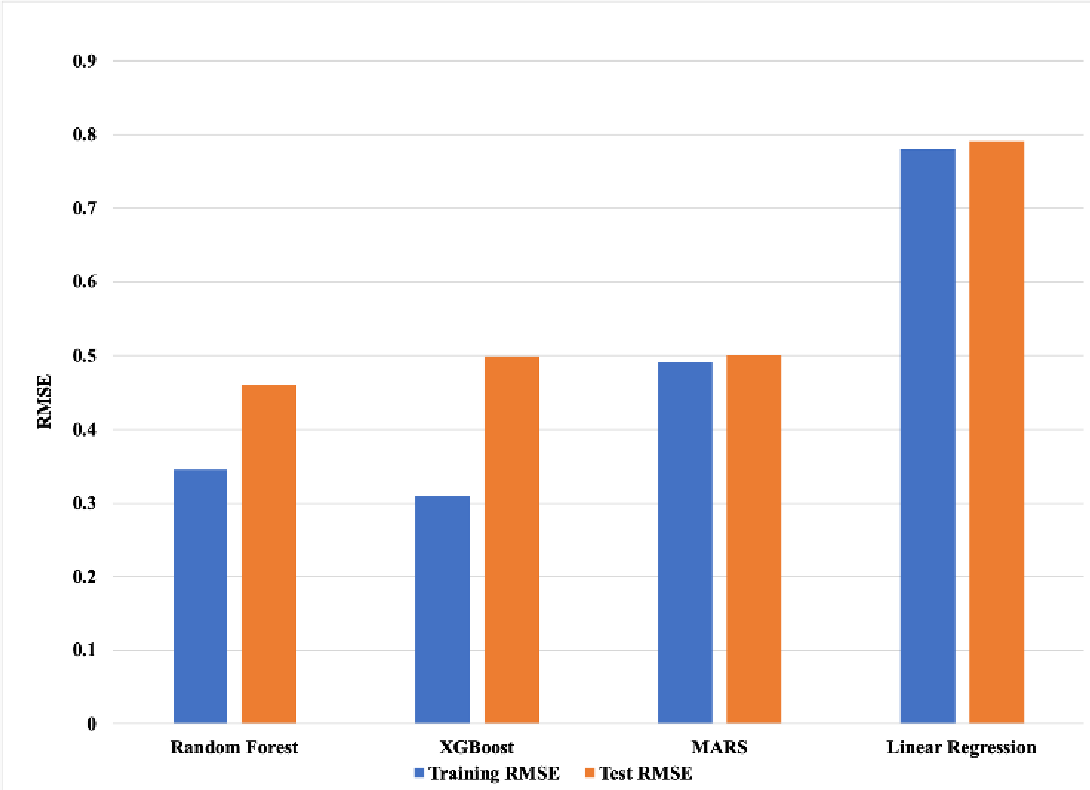


Fig. 4: Comparison of model RMSE values

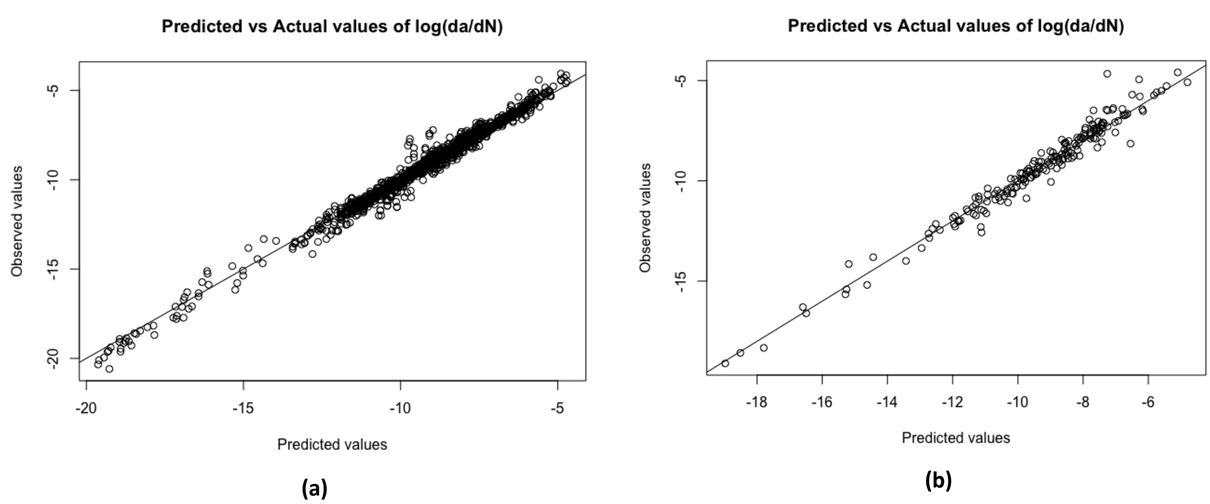


Fig. 5: Comparison of observed and predicted transformed da/dN using RF model: a) Training data, and b) Test data